

Homework 4: Comparing Generative Sequential Models of the Bee Movie

Daniel Dinh

Introduction

In machine learning, sequential models, such as recurrent neural networks (RNN), offer an alternative way to predict data sequences. Unlike traditional model, sequential models are capable of capturing patterns and the context of data making them capable of handling data in which the position and order of the data matters. More specifically these types of models can be used to handle either natural language or be used to handle time series analysis. Within this project we will be discussing how these sequential models can be used in tandem with a generative model. Generative models allow for the creation of entirely new data instances. With this project specifically we will be using two sequential models followed by a basic generative model to be able to create new text.

The significance of understanding these kinds of models is due to the increasing popularity and usage of AI and chatbots such as ChatGPT. Since these large language models primary use both sequential and generative models it is important to understand how these models to have a better understand of some of the most popular technology at the moment.

The dataset used was twofold with Project Gutenberg being used for the first dataset. The first sample text was a copu of “The Fall of the House of Usher” by Edgar Allen Poe. This book was primarily used for the initial building and testing of the model. Afterwards the second and final text that was used is the script of the 2009 Bee Movie which was written by Jerry Seinfeld.

The objective of this project is to build two slightly different generative models, both based on recurrent neural networks. These models undergo training on the same sample text, differing only in specific architectural aspects. Subsequently, the trained models are put to the test, generating new text based on random samples extracted from the training data. Through this process, the accuracies and the output of the generated model is analyzed to see how these models perform with the given texts.

In the following sections, I will delve into the analysis of the chosen text data, elaborate on the model architectures and training methodologies employed, present the results of the generated

sequences, and conclude with reflections on the overall process. This report aims to offer a comprehensive understanding of the experiment, catering to both technical and non-technical audiences alike.

Analysis

Within this section, the data used will be covered more in detail as well as what steps were taken to be able to use the data.

As mentioned before, there were primarily two texts used: one being “The Fall of the House of Usher” by Poe and the other being the Bee Movie script by Seinfeld. With both texts several steps were taken to be able to have the texts in a usable state. First, each text was saved as a plain text file, and was loaded into Python using the default open, read, and close methods built into Python’s libraries. We then take the files and clean them to first standardize all of the words so that they are all lowercase. In addition to this we also remove all punctuation such that we are not having punctuations as tokens within our model as well.

After cleaning through our text files we are able to measure the number of words within our model’s vocabulary, in this case the vocabulary are all of the words that appear uniquely within the each text file which we call tokens. The number of unique tokens for House of Usher was ~2500 and for the Bee Movie the number of unique tokens was ~1800. What these tokens represent are the vocabulary or dictionary that the model associates with the given input text to be trained upon. From there the tokens are saved in 100 token long sequences and saved to another file with all of the combinations of 100 word long sequences that appear within the given input text. From here both the input text has been formatted and saved in a way that is usable for both the sequential and generative models.

Methods

At the beginning of this assignment examples of sequential and generative models were provided by Dr. Parlett and were to serve as the baseline for the assignment. Much of the data cleaning and tokenization for the sequential model was provided from this example as well as the basis for the generative model. It was primarily within the sequential model that we were given the freedom to alter and adjust the models to fit our needs and try to increase the efficiency as much as possible.

To begin the building of the models, the types of sequential models were put to consideration. With the knowledge that two models had to be created, it was decided upon using one model as a GRU sequential model and the other as a LSTM model. These models both work somewhat similarly however the main difference between the two is the addition of the previous cell state as an input within the LSTM model. What this allows is for the creation of two somewhat

similar models with the exception of the what type of sequential model was used. In addition other changes between the model was the output shape of the initial embedding layer and the addition of another dense layer within the LSTM model.

Within the GRU model we start off by taking a single sequence in this case 100 words and split the X as all words except for the last one which is set to Y. From here we build a GRU model by first going through a embedding layer with an input of the entire vocabulary of our text, output size of 50, and an input sequence length of 100 which we pre-define. From here we go through two GRU layers with a output size of 100 and return sequence set to True. From here we go to one final GRU layer that does not return sequences but still has 100 as it's output size and a dropout value of 0.2 for some regularization. Afterwards we then go through 3 Dense layers with a dropout layer set between the 2nd and the 3rd. The first two layers are identical with both being set to output size of 100 and an activation function of relu. Afterwards we have a standard dropout layer as mentioned previously before entering into the final dense layer. In this layer we are having the output size be the entire vocabulary size and the activation function as softmax. What this layer will do is output the probability of the last token being a certain word within the vocab. This is what we use to predict what the next word should be as we simply pick the word that has the highest probability according to our model as the next word. Below is the summary of the GRU model:

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 100, 50)	91900
gru_3 (GRU)	(None, 100, 100)	45600
gru_4 (GRU)	(None, 100, 100)	60600
gru_5 (GRU)	(None, 100)	60600
dense_2 (Dense)	(None, 100)	10100
dense_3 (Dense)	(None, 100)	10100
dense_4 (Dense)	(None, 100)	10100
dense_5 (Dense)	(None, 1838)	185638
Total params: 474638 (1.81 MB)		
Trainable params: 474638 (1.81 MB)		
Non-trainable params: 0 (0.00 Byte)		

Figure 1: Screenshot of the GRU Model Summary

With the LSTM model we see somewhat of a similar model minus the few changes mentioned before. As stated the output shape for the LSTM model is set to 100 and not 50. From there we have 2 LSTM layers that are set similar to output shape of 100 and returning the sequences. Once again the final LSTM layer is similar to the GRU one with output shape of 100 and a dropout rate of 0.2. From here we go through only 1 Dense layer with relu activation function after which a dropout layer of 0.2 is added followed by the final dense layer which is similar to the GRU model and serves as the layer in which we make our prediction of what the next word is. Below you can see the screenshot of the LSTM model summary:

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 100, 100)	183800
lstm (LSTM)	(None, 100, 100)	80400
lstm_1 (LSTM)	(None, 100, 100)	80400
lstm_2 (LSTM)	(None, 100)	80400
dense_6 (Dense)	(None, 100)	10100
dense_7 (Dense)	(None, 1838)	185638
Total params: 620738 (2.37 MB)		
Trainable params: 620738 (2.37 MB)		
Non-trainable params: 0 (0.00 Byte)		

Figure 2: Screenshot of the GRU Model Summary

From here both models were individually fitted and trained for initially 100 epochs where the accuracy was tracked. Initially the model was only trained for 100 epochs and the initial accuracy recorded with the accuracy of the GRU being ~65% and the LSTM accuracy ~23.5%. After talking with Dr. Parlett, it was decided that the models should continue training until the accuracies were at a satisfactory level. Hence the number of epochs was increased and training was continued until the final accuracies was ~65% for both the GRU and the LSTM model. The amount of epochs seen for both are what the final amount required to be able to get the accuracy to this level for both models.

Once the models were built to a satisfactory level then the generative model was implemented by taking a random sequence of tokens from the sequence document and then looped through applying the trained model to the sequence of words to generate 100 new words that would be outputted to a file. For example the last 99 words from the seed text was taken to predict the 1st new generated word, then the last 98 words plus the recently generated word was taken to generate the 2nd new word and so on. Once this process is repeated until there are 100 new generated words after which the new text is saved to a Python and this is repeated 10 times to generate 10 new 100 word sequences. From here the general analysis of the performances of the models are analyzed by examining the accuracies and the generated text of the models.

Results

Within this assignment, the main metric used was the accuracies of the models that were found when training and the generated text that was saved within the output files. For both models they were trained and built until the end accuracy of the models was ~65% for both the GRU and the LSTM model. What was interesting was how many epochs were required to reach this point, as for the GRU only 250 epochs was required while the LSTM needed 300 epochs. Moreover, there were times within training where the model would not train at all, times when the model would overfit, and others when the model would train normally. This is

why it was stated that the average accuracies for the models as the performance of the models would fluctuate so much between each time the code was run. Within both models we can see that there are areas of text that make sense and other areas that are nonsense. For example within the GRU generated text we see:

```
...just heard em bear week next week theyre scary hairy and here live...
```

which does semantically make somewhat sense about how bears are hairy and scary and live in the area but syntatically does not make any sense. This is most likely from the accuracy of the model not being the best but still able to discern generally what the meaning of the text. There are also instances where the code just generates code that is repeated before finally able to break through and generate something new as seen here also in the GRU output file:

```
...you snap out of it you snap out of it you snap out of it you snap out  
of it hold it why oome on its my turn or a bees hello...
```

As you see the phrase “you snap out of it” is repeated multiple times before the model is able to generate something new. This is a sign of our model’s performance as on one hand we are able to generate english phrases that make sense but it is repeated so the entire protion of the sequence does not make sense.

These patterns also appear within the LSTM model but it seems that the model did not perform as well as the GRU model as it appears that several of the generated code does not make sense. For example:

```
...also wow all example you grab my smell of flames not as thinking bee fellow move ...  
...were lucky well know wed not like to dont say something now our  
wings i are as the steps into you go ....  
...what i have all for i make to go to to to go up youre highfive  
were first with the humans ...
```

As seen above, the lines of generated text tend to not make as much sense as the GRU model. The LSTM model appears to be able to generate more words and phrases that don’t make sense not only syntatically but also semantically. So it’s interesting to note that the two models although appearing to have the same accuracy to still perform and generate texts at two different levels of accuracy.

GRU Generated text

Below are all of the generated sequences of the GRU model:

1. to drain the old stinger to oheck of nature he why you got i got just about afraid it with your little mind games whats some italian vogue mamma mia thats a lot of pages a lot

of ads remember you dont flayman flying here your life more valuable than mine funny i just have to flower how pea you believe something stinks in here the love the smell of flowers how do you like the smell of flames as to only is it not taking sides stealing its its hey afraid they know what only wanted what make on

2. been dont know if you know anything about fashion of this does from all really now we tie we like the clients a could never deal this they smoke long seem you could have the cicada with the was just is this honey the say out i know you said guatemalan these told the marry a watermelon is that a was joke of the kind of stuff we do yeah different so what are you gonna want barry what work i know i know i were that my only is the hive but i pollen bees are it were going
3. more honey for us oool im picking up a lot of bright yellow oould be daisies be funny a needs they remember its its one fun we wont rest of you have the clients youre dead benson me the moving flower affirmative that was the flower on this is the coolest what is it i dont know but im loving some color it smells good not like this is but i just are i know at the know to smell an have my pinhead deli you little heard your flayman myself yeah those balloon little be i princess what honesco
4. just what bees know the dont dont know a lot of the bees of to human disaster up gusty hospitals of hello weekend we have and its all ours you heard you and of to flower i know to the bottom of all of a lot i of autograph dying it the perfect cant work its i humans one guess or a had you have to pea we could no all well no one around a gotta stinger crashing know i suppose crazy working so getting guy i dont talk to was this stinger their way where you getting right
5. got giant wings huge engines i cant fly a plane why not isnt john travolta a pilot yes how hard could a be really everything against an need some lightning this is bob bumble we have some latebreaking news from jfk airport where a suspenseful scene is developing barry benson fresh from his legal victory it like is dead a appropriate it tail casually through would out a nickel beemen listen crew flowers we have a storm in a area and two individuals at the controls with absolutely no flight experience just a minute theres a bee and the fault
6. ohemicaly oareful guys its a little grabby my sweet lord of have a dont heard the anchor guys to havent sting the stirrer what man and who like a crud of a remember out of position rookie ooming in at you like a missile help me i dont think you will check barry this going you know im are skills dont dont know every belong to a know it it bad scared i only cant meant to you flower a got the know off if you something was driving always i i have it giant bug flower a guess too
7. bats also i got a couple of reports of root beer being poured on then sorry in the think me i bee soothing he was behind thats to no ok it all flowers this theyve to number going at our talking to bees absolutely out you hexagon buzz im hate buzz buzz buzz buzz buzz buzz buzz buzz black and yellow hello hello hello nectar is order like flowers nectar pollen you be sting produce things check oome in i striped im kidding out you cant got to trade with jewish ladies lets move it out pound those buddy
8. you snap out of it you snap out of it you snap out of it you snap out of it you snap out of it you snap out of it you snap out of it hold it why oome on its my turn or a bees

hello all i dont know hello benson i pick this time its out of this pollen here there queens out there if by a human hives and no behind all right lets drop this tin cant have what yet out i do maybe stinks you have wait its oome in you cant see maybe

9. stinger flowers pollen they can have with all ok ladies lets move it out pound those petunias you striped stemsuckers all of you drain those flowers wow im out i cant believe im out so blue i feel so fast and free box kite wow flowers this is blue leader we have roses visual bring it around degrees and hold roses degrees roger bringing it around stand to the side kid its go was to bee flower no another bee know we know im see pollination up close no sir i pick up some way of weather out a i
10. the bottom from a guest even though you just heard em bear week next week theyre scary hairy and here live always leans forward pointy shoulders squinty eyes very jewish in tennis its as you make will is ive it was my grandmother ken shes so her backhands a joke to are you take advantage of that quiet please actual work going on here i be so same bee yes it are im helping him sue the human race hello hello a pea you will smoker i start this candy oant returning on we krelman dated any i believe you

LSTM Generated text

Below are generated text for the LSTM model:

1. favor of the bees vanessa we won you knew you go do a little scary term on is they have all before i think to also bred will yet another example of same only because im not an bee or lets step a privilege and the anchor whassup in bees when one of what also wow all example you grab my smell of flames not as thinking bee fellow move it a york ill see what is his only because what happened well happened its barry ok he doesnt beekeepers these flowers glasses quotes on the bottom from the guest
2. er in thank you thank you i see from you gotta get home like a row of it jars do far doing this daisies them i dont know bees not we never do vacations boy not i fly sure thats pollen what is we have it over dont worry the only thing i have to do to turn this is around theres you know a moment are you was hey was were bees a trap of course im helping to do with the topsyturvy world long time what pick you helping me the ladies see here its like well all
3. got you have to do a terrific job out is me a bee chance me me its will want to say we just go youre like a bit of pomp under you thought it you only think bees all as a row of honey jars as far as a tournament and honesco in this vanessa roses thinking you see a crumb well do it our job thats im hello of what kind the marry i get kenny when that a pointed krelman day some make you know that the sleeves and suspenders im has nauseous him up all weve got 4.dont want to put you out its no is here they are who just what im only asked fly down bad stirrer im live take right away tonight up hover with session we have it thats what the time what have you have to the pea in you met im cinnamon and theres right honeyburton what going a little weird its he hello me thats

- i had id go set its barely we wants a little weird thats the bees but a great take it me
 profiting hi wait im sorry to one here right jocks im time is theres
4. all i got to have to carob pea to live it could you sting how you do him for your little
 queen oh honey attempting you poor florist you matters adam oongratulations i stinger
 mr liotta pack belated stink check oant bear what you awkward three wings college in
 that in the bees died like the real name and real remove you could be careful hard like
 it mosquito the species player im going a bee have a miss vanessa least in me our arent
 where the man where trouble the honey show im not a moment do you got
 5. dont see what its work all i dont know you need to like now wasnt you did you where to
 go to the last chance you was out with a bee do you ever have go right we never say like
 i know it they intend you dont know are i need to all like it now they have its out its a
 last humans there not like it right our homes just be mamas right with to like to could
 heard it mr ladies barry italian six miles is no pasadena bloome youre honex our barry
 thats npr attorney
 6. hive you did come back different some theres degrees talking you start winter into a little
 life of why other talking next were yelling youre all with i dont know date i know who
 fly its all orientation person to just mixed for the game slogan im much to like this like i
 dont know its its capable you know i want it bees the little bit of destruction and vanessa
 roses are thinking bee surprised we about hes what i have all for i make to go to to to
 go up youre highfive were first with the humans
 7. i got to have to carob pea to live it could you sting how you do him for your little queen
 oh honey attempting you poor florist you matters adam oongratulations i stinger mr
 liotta pack belated stink check oant bear what you awkward three wings college in that
 in the bees died like the real name and real remove you could be careful hard like it
 mosquito the species player im going a bee have a miss vanessa least in me our arent
 where the man where trouble the honey show im not a moment do you got a
 8. benson from to learn it all i the little perfectly games whats you know barry barry hell
 have never been like the entire animal name on bees it it like wow is we got a cicada
 from awful im we doesnt experience just a minute that goes under i can got a instinct
 sure my little face thats your stinger janet you just been one we do it all not got i dont
 know is its if why like i know what its pollen talking to me is a bee thing but but im got
 to get me wait i
 9. were lucky well know wed not like to dont say something now our wings i are as the steps
 into you go so if thats around i slow into well that a minute they have to sure to want
 to go bred for with what just who dont dont dont work you leave you call home all in a
 lot of ads remember the nose down flowers not longer doing to make this the ladies see
 here its like me if he just all for i make a couple help more its doing we dont get but he
 were so

Reflection

Within this assignment, I was able to learn more about the sequential models and the necessary tips and tricks to try and get the model as effective as possible. I was able to further understand the metrics of the sequential model specifically if there was a baseline for accuracy and how to train it in a way such that the model still performed well while being cheap on resources such as RAM/GPU and time. Something I also learned was how to finally get the files and code running within the server as with prior assignments I was able to work from just my Google Colab but with this it was a necessity to have the model trained and tested on the server so this assignment was able to finally allow me the ability to work on the server.

Moving forward I would most likely try to go to Dr. Parlett's Office Hours more as there were times when I needed help and needed to clarify but would ask a classmate instead of Dr. Parlett. Going to her OH would also help me find out information that I would've needed to check if I had gone earlier. An example being the need to train my models further instead of just sticking to the accuracies I got after running the models after 100 epochs. It was only after asking Dr. Parlett and telling her my accuracies that I was told to try and retrain my models to get the LSTM accuracy higher which was done in the end but necessitated another late day to be used.